Biometrics 64, 1–22 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Integration of Survival Data from Multiple Studies

Steffen Ventz[1,*], Rahul Mazumder[2], and Lorenzo Trippa[3]

[1]Department of Data Statistics, Dana-Farber Cancer Institute and

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A

[2]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.

email: steffen.ventz.81@gmail.com

Summary: We introduce a statistical procedure that integrates datasets from multiple biomedical studies to predict patients' survival, based on individual clinical and genomic profiles. The proposed procedure accounts for potential differences in the relation between predictors and outcomes across studies, due to distinct patient populations, treatments and technologies to measure outcomes and biomarkers. These differences are modeled explicitly with study-specific parameters. We use hierarchical regularization to shrink the study-specific parameters towards each other and to borrow information across studies. The estimation of the study-specific parameters utilizes a similarity matrix, which summarizes differences and similarities of the relations between covariates and outcomes across studies. We illustrate the method in a simulation study and using a collection of gene expression datasets in ovarian cancer. We show that the proposed model increases the accuracy of survival predictions compared to alternative meta-analytic methods.

Key words: Hierarchical Regularization, Meta-Analysis, Penalized Regression, Risk Prediction, Survival Analysis.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13517

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Biomedical technologies enable the use of omics information for prognostic purposes, to quantify the risk of diseases or to predict response to treatments. Risk stratification in oncology often utilizes a set of biomarkers to predict cancer progression or death within a time period. The number of covariates can exceed the sample size. This makes the identification of relevant genomic features for risk prediction and the development of accurate models challenging. Penalized regression and methods that utilize multiple datasets have been discussed in this context. Regularization methods enable parameter estimation and prediction when the number of predictors is large (Tibshirani, 1997). Meta-analyses (DerSimonian and Laird, 1986) and integrated analyses (Conlon et al., 2006) combine information from multiple studies for parameter estimation and prediction (Hedges and Olkin, 2014). These statistical procedures improve the estimation of parameters of interest with respect to single-study estimates if the covariate-outcome relations are similar across studies (Bernau et al., 2014; Waldron et al., 2014). For instance (Bernau et al., 2014; Waldron et al., 2014) showed that meta-analytic procedures tend to outperform the prediction accuracy of models developed using only a single study. But (Waldron et al., 2014; Trippa et al., 2015) also discussed variations of the covariate effects across cancer studies, due to differences in assays, treatments and patient populations.

We introduce a model for the integrated analysis of a collection of datasets, with the aim of improving the prediction accuracy compared to single-study models and meta-analytic procedures. We use study-specific parameters in covariate-outcome regression models. These parameters are estimated borrowing information across studies with hierarchical regularization, which shrinks the study-specific regression coefficients towards each other. We use a $K \times K$ similarity matrix representative of differences and similarities of the covariate effects

across K studies. The regression parameters of each study are shrunken more towards the parameters of similar studies and less towards the remaining studies.

In previous work on integrative analyses Liu et al. (2011) discussed Bayesian methods and variable selection for accelerated failure time models. Hierarchical models for multi-study gene expression analyses have been developed in (Conlon et al., 2006, 2009, 2012), and (Ma et al., 2011) studied penalized regression methods for integrative analyses, focusing on binary outcomes and accelerated failure time models. For case-control analyses with multiple datasets (Liu et al., 2013) proposed an adaptive group-LASSO procedure, and Cheng et al. (2015) extended the approach using different regularization techniques. In the following sections we introduce a procedure which builds on the work that we mentioned. The procedure shrink study-specific parameters towards each other accounting for the degrees of similarity specific of each pair of studies.

2. The model

We consider K studies with time-to-event outcomes and predictors such as gene expression measurements. For each study $k = 1, \cdots, K$, $\boldsymbol{Y}_k = \{Y_{k,i}\}_{i=1}^{n_k}$ indicates (possibly censored) survival times of n_k individuals and $\boldsymbol{C}_k = \{C_{k,i}\}_{i=1}^{n_k}$ denotes the vector of censoring variables, where $C_{k,i} = 1$ if $Y_{k,i}$ is an observed event time and it is zero if there is censoring at time $Y_{k,i}$. The vector $\boldsymbol{x}_{k,i} \in \mathbb{R}^p$ represents a set of p predictors and $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n_k})'$. Lastly, $\mathcal{D}_k = (\boldsymbol{Y}_k, \boldsymbol{C}_k, \boldsymbol{X}_k)$ indicates the data from study k and $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$.

The method that we discuss allows users to predict patient survival (i) in a population k of interest (e.g. patients diagnosed at a specific institutions k) with samples $\mathcal{D}_k$ from the population k available in the data collection $\mathcal{D}$, or (ii) in other populations, distinct from the K training datasets (e.g. patients diagnosed at institutions not represented in $\mathcal{D}$).

We assume that failure times in each study k follow a proportional hazards model (Cox, 1972) with baseline survival function $S_k(\cdot)$ and study-specific coefficients $\boldsymbol{\beta}_k \in \mathbb{R}^p$. The

approach that we will describe can be applied to alternative time-to-event models, for instance to accelerated failure time models (Wei, 1992), computations would only require minor modifications.

Inference is based on Breslow's modification of the partial log-likelihood functions (Cox, 1972) for (possibly tied) survival times,

$$l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) = \sum_{\ell=1}^{m_k} \left\{ \tilde{\boldsymbol{x}}'_{k,\ell}\boldsymbol{\beta}_k - d_{k,\ell} \log\left(\sum_{i:Y_{k,i}\geqslant t_{k,\ell}} \exp\{\boldsymbol{x}'_{k,i}\boldsymbol{\beta}_k\} \right) \right\},$$

where $\left\{t_{k,\ell}\right\}_{\ell=1}^{m_k}$ are the m_k unique event times in study k, $d_{k,\ell}$ denotes the number of observed events at time $t_{k,\ell}$, $\tilde{\boldsymbol{x}}_{k,\ell} = \sum_i \boldsymbol{x}_{k,i} I(C_{k,i} = 1, Y_{k,i} = t_{k,\ell})$, for $\ell = 1, \ldots, m_k$, and $I(A)$ is the indicator function of the event A. When the number of predictors exceeds the number of observed events a unique maximum partial-likelihood estimate does not exist and maximization of the regularized likelihood function $l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) - R\left(\boldsymbol{\beta}_k\right)$ has been proposed (Tibshirani, 1997) to estimate the covariate effects $\boldsymbol{\beta}_k$. Popular approaches include the LASSO, ridge, elastic-net and the bridge penalties to name a few (Hoerl and Kennard, 1970; Tibshirani, 1997; Fu, 1998).

We introduce a model with study-specific parameters $\boldsymbol{\beta}_k$, and a latent parameter $\boldsymbol{\beta}_0$, which can be interpreted as the mean parameter across studies. Some studies will have similar vectors $\boldsymbol{\beta}_k$ due to similarities in the assays and patient populations, while other studies might be considerably different (Bernau et al., 2014; Waldron et al., 2014). We estimate the vector $\boldsymbol{\beta}_k$ by borrowing information from studies $k' \neq k$ that are similar to study k. At the same time, studies k' that differ substantially from study k will have little influence on the estimation of $\boldsymbol{\beta}_k$. The parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_K)$ are estimated using the regularized likelihood

$$l_R\left(\boldsymbol{\beta}\right) = \sum_{k=1}^{K} l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) - R_0(\boldsymbol{\beta}_0) - R_1\left(\boldsymbol{\beta}\right). \tag{1}$$

Here the parameters $\boldsymbol{\beta}_k$ can be interpreted as a noisy realization of $\boldsymbol{\beta}_0$, the average effect across studies. The non-negative function $R_0(\cdot)$ regularizes $\boldsymbol{\beta}_0$ and is zero when $\boldsymbol{\beta}_0 = \mathbf{0}$ (for example a ridge penalty). Similarly, the non-negative function $R_1(\cdot)$ is zero when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$ (see below for examples) and is used to borrow information across studies in the estimation of $\boldsymbol{\beta}$. In our applications below we will use $\widehat{\boldsymbol{\beta}}_0$ for risk predictions of patients in populations $k > K$ that are not represented in our collection of K studies, whereas for patients belonging to populations $k = 1, \ldots, K$, the estimate $\widehat{\boldsymbol{\beta}}_k$ can be directly used for risk predictions.

Penalized maximum likelihood estimates based on (1) have a Bayesian interpretation. Consider a Bayesian model for the unknown parameters $\boldsymbol{\beta}$, with prior probability $Pr(\boldsymbol{\beta}_0) \propto e^{-R_0(\boldsymbol{\beta}_0)}$ for the vector $\boldsymbol{\beta}_0$ and $Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \boldsymbol{\beta}_0) \propto e^{-R_1(\boldsymbol{\beta})}$ for the study specific parameters conditionally on $\boldsymbol{\beta}_0$. The approximate posterior density of $\boldsymbol{\beta}$ includes the the partial likelihood (see Sinha et al. (2003) for a formal justification) and is proportional to

$$Pr_{PL}\left(\boldsymbol{\beta}|\mathcal{D}\right) \propto Pr(\boldsymbol{\beta}_0) Pr(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K | \boldsymbol{\beta}_0) \prod_{k=1}^{K} e^{l(\boldsymbol{\beta}_k, \mathcal{D}_k)}. \tag{2}$$

Therefore the mode of (2) coincides with the parameter $\boldsymbol{\beta}$ that maximizes (1). If we set $R_1\left(\boldsymbol{\beta}\right) = \sum_k \tilde{R}_1\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0\right)$ with $\tilde{R}_1\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0\right) \geqslant 0$, the Bayesian model (2) incorporates the assumption that the covariate effects are, *a priori*, exchangeable and have, conditionally on $\boldsymbol{\beta}_0$, independent and identically distributed effects $\boldsymbol{\beta}_k$. For example $\tilde{R}_1\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0\right) = ||\boldsymbol{\beta}_k - \boldsymbol{\beta}_0||_2^2/(2\nu_1)$ and $R_0\left(\boldsymbol{\beta}_0\right) = ||\boldsymbol{\beta}_0||_2^2/(2\nu_0)$, $\nu_1, \nu_0 > 0$ is consistent with the commonly utilized hierarchical normal prior model with, *a priori*, correlations $\mathrm{Cor}(\beta_{k,j}, \beta_{k',j}) = \nu_0/(\nu_1 + \nu_0) > 0$ for $k' \neq k$. This regularization implies positive and symmetric borrowing of information for all pairs $k \neq k'$ of studies, and may not be appropriate for groups of studies with different patient populations.

For the parameter $\boldsymbol{\beta}_0$ we use the elastic-net penalty,

$$R_0(\boldsymbol{\beta}_0) = \lambda_0||\boldsymbol{\beta}_0||_1 + \frac{\lambda_1}{2}||\boldsymbol{\beta}_0||_2^2, \tag{3}$$

$\lambda_0, \lambda_1 \geqslant 0$, with LASSO and ridge penalty as special cases, when $\lambda_1 = 0$ and $\lambda_0 = 0$.

To account for differences and similarities of the available studies, we use

$$R_1\left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} \frac{1}{2}||\boldsymbol{\beta}_{1:K,j} - \beta_{0,j}\mathbf{1}||_{\boldsymbol{\Sigma}}^a \tag{4}$$

in (1), where $\mathbf{1}$ is a K-dimensional vector with one on each component, $\boldsymbol{\beta}_{1:K,j} = (\beta_{1,j}, \cdots, \beta_{K,j})'$ refers to the j-th covariate, and $||\boldsymbol{x}||_{\boldsymbol{\Sigma}} = \sqrt{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$. The symmetric matrix $\boldsymbol{\Sigma}$ is positive-semidefinite and enables differential borrowing of information across studies.

For $a = 2$, the maximizer of (1) is equivalent to the posterior mode when, *a priori*, the coefficients $\boldsymbol{\beta}_{1:K,j}, j = 1, \ldots, p$, across studies are modeled (conditionally on $\boldsymbol{\beta}_0$) with a multivariate normal distribution with mean $\beta_{0,j}\mathbf{1}$ and covariance matrix $\boldsymbol{\Sigma}$. In this case $\Sigma_{k,k'} = 0$ implies that, conditionally on $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k'}$ are, *a priori*, independent. Whereas a covariance $\Sigma_{k,k'} > 0$ indicates similarities between $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k'}$.

For $a = 1$, $R_1(\boldsymbol{\beta})$ in (4) is proportional to the sum of p Mahalanobis distances, between the $\boldsymbol{\beta}_{1:K,j}$ coefficients and the K-dimensional vectors $\beta_{0,j}\mathbf{1}$, with covariance matrix $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma} \propto \boldsymbol{I}$ this penalty reduces to the group LASSO (Cheng et al., 2015; Liu et al., 2013), with one group for each $j = 1, \cdots, p$.

The regularization parameters λ_0, λ_1, a, and $\boldsymbol{\Sigma}$ determine (i) the similarity of the estimates $\widehat{\beta}_{1,j}, \cdots, \widehat{\beta}_{K,j}$ across studies, including the number of identical study-specific estimates $\widehat{\beta}_{k,j} = \widehat{\beta}_{k',j}$ and (ii) the sparsity of $\widehat{\boldsymbol{\beta}}_0$ (the number of components $\widehat{\beta}_{0,j} = 0$).

For $\lambda_0, \lambda_1 \geqslant 0$, and $a = 1$, the choice of $\boldsymbol{\Sigma}$ can lead to identical study specific estimates $\widehat{\beta}_{1,j} = \cdots = \widehat{\beta}_{K,j}$. We provide an example with $\lambda_0 = 0$ and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{z}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}_0, k = 1, \cdots, K$, and define $h(\boldsymbol{z}) = \max_{\boldsymbol{\beta}_0}\left\{\sum_{k=1}^{K} l\left(\boldsymbol{z}_k + \boldsymbol{\beta}_0, \mathcal{D}_k\right) - R_0\left(\boldsymbol{\beta}_0\right)\right\}$. The map $h(\boldsymbol{z}) - \sum_{1 \leqslant j \leqslant p} ||(z_{1,j}, \ldots, z_{K,j})||_{\boldsymbol{\Sigma}}^a$ bounds the re-parametrized regularized log-partial-

likelihood ($l_R^* : [\boldsymbol{\beta}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K] \to \mathbb{R}$). If we specify $1/\sigma > \max_{j,k} |\partial 2h(\boldsymbol{z})/\partial z_{k,j}|$ at $\boldsymbol{z} = \mathbf{0}$, then the concave function (1) is maximized at $\widehat{\boldsymbol{\beta}_1} = \cdots = \widehat{\boldsymbol{\beta}_K} = \widehat{\boldsymbol{\beta}_0}$. More generally, if we don't assume a diagonal $\boldsymbol{\Sigma}$ and indicate with σ^2 the largest eigenvalue of $\boldsymbol{\Sigma}$, then the equalities $\widehat{\boldsymbol{\beta}_k} = \widehat{\boldsymbol{\beta}_0}$ hold when $1/\sigma > \max_{j,k} |\partial 2h(\boldsymbol{z})/\partial z_{k,j}|$ at $\boldsymbol{z} = \mathbf{0}$.

When $a \geqslant 1$, $\boldsymbol{\Sigma}$ is positive-definite and $\lambda_0 > 0$ or $\lambda_1 > 0$, the regularized log-partial-likelihood (1) is concave. If we fix $\lambda_1 \geqslant 0$, $a \geqslant 1$ and the positive-definite matrix $\boldsymbol{\Sigma}$, then the number of components $\widehat{\beta}_{0,j}$ equal to 0 increases with λ_0. For instance, consider $a > 1$ in (1), $\lambda_1 = 0$, and $g(\boldsymbol{\beta}_0) = \max_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K} \left\{ \sum_{k=1}^K l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) - R_1\left(\boldsymbol{\beta}\right) \right\}$. The concave map $g(\boldsymbol{\beta}_0) - \lambda_0 ||\boldsymbol{\beta}_0||_1$ bounds the regularized log-partial-likelihood. If we choose λ_0 larger than $\max_{1 \leqslant j \leqslant p} |\partial g(\boldsymbol{\beta}_0)/\partial \beta_{0,j}|$ at $\boldsymbol{\beta}_0 = \mathbf{0}$, then (1) is maximized at $\widehat{\boldsymbol{\beta}_0} = \mathbf{0}$. In contrast, for values λ_0 below this maximum some of the estimates $\widehat{\beta}_{0,j}$ will be different from zero.

In several applications, where the primary aim is prediction, sparsity in the estimation of $\boldsymbol{\beta}_0$ is not necessary and one may set $\lambda_0 = 0$. In other settings it is important to identify a practical and parsimonious set of predictors that guarantees prediction accuracy in support of clinical decisions Feng et al. (2021). The regularization function $R_0(\cdot)$ for $\boldsymbol{\beta}_0$ allows to opt for a sparse estimate of $\boldsymbol{\beta}_0$ ($\lambda_0 > 0$) or not ($\lambda_0 = 0$).

3. Parameter estimation

We use the alternating direction method of multipliers (Boyd et al., 2011) to estimate $\boldsymbol{\beta}$. We first formulate the optimization of (1) with respect to $\boldsymbol{\beta}$ as a constrained convex minimization problem

$$\min_{(\boldsymbol{\beta},\boldsymbol{z})} \left\{ \sum_k -l(\boldsymbol{\beta}_k, \mathcal{D}_k) + R_0(\boldsymbol{\beta}_0) + R_1(\boldsymbol{z}) \right\},$$

where $\boldsymbol{z} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_K)', \boldsymbol{z}_k \in \mathbb{R}^p$, subjected to the affine constraints $\boldsymbol{\beta}_k = \boldsymbol{z}_k, k = 0, \ldots, K$. We then introduce for this minimization problem the scaled augmented Lagrangian

$$L_\rho(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{u}) = \sum_{k=1}^{K} -l(\boldsymbol{\beta}_k, \mathcal{D}_k) + R_0(\boldsymbol{\beta}_0) + R_1(\boldsymbol{z}) + \sum_{k=0}^{K} \frac{\rho}{2} ||\boldsymbol{\beta}_k - \boldsymbol{z}_k + \boldsymbol{u}_k||_2^2, \tag{5}$$

where $\rho > 0$, with $\boldsymbol{u} = (\boldsymbol{u}_0, \ldots, \boldsymbol{u}_K), \boldsymbol{u}_k \in \mathbb{R}^p$. For a fixed $\rho > 0$, the algorithm that we describe converges to a solution $\boldsymbol{\beta} - \boldsymbol{z} = \mathbf{0}$ that maximizes (1). The algorithm minimizes (5) iteratively (i) with respect to $\boldsymbol{\beta}$, and (ii) with respect to $\boldsymbol{z}$, and (iii) then it updates $\boldsymbol{u}$ to $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{\beta} - \boldsymbol{z}$, while keeping at each of the three steps the remaining two parameters fixed. At each iteration of the algorithm the minimization of (5) with respect to $\boldsymbol{\beta}$ (*step i*) can be carried out independently for each component $\boldsymbol{\beta}_k$, $k = 0, \ldots, K$, and the minimization with respect to $\boldsymbol{z}$ (*step ii*) can be carried out independently for each $\boldsymbol{z}$ component $j = 1, \cdots, p$.

The algorithm starts with an initial estimate of $\boldsymbol{\beta}$ (we use $\mathbf{0}$ or preliminary estimates of $\boldsymbol{\beta}_k, k = 0, \cdots, K$), $\boldsymbol{\beta} = \boldsymbol{z}$ and $\boldsymbol{u} = \mathbf{0}$. At each iteration, in *step i* the algorithm minimizes (5) with respect to $\boldsymbol{\beta}$, keeping $\boldsymbol{z}$ and $\boldsymbol{u}$ fixed, by setting $\boldsymbol{\beta}_0 = \dfrac{S\big(\rho(\boldsymbol{z}_0 - \boldsymbol{u}_0), \lambda_0\big)}{\rho + 2\lambda_1}$ where $S(\boldsymbol{x}, \lambda)$ is the coordinate-wise soft-thresholding function $s(x_j, \lambda) = (1 - \lambda/|x_j|)_+ x_j$, and

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{b}} \Big(- l(\boldsymbol{b}, \mathcal{D}_k) + \rho ||\boldsymbol{b} - \boldsymbol{z}_k + \boldsymbol{u}_k||_2^2/2 \Big)$$

for $k = 1, \cdots, K$. We used a quasi-Newton algorithm (Byrd et al., 1995) for the latter minimization.

In *step ii*, the algorithm minimizes (5) with respect to $\boldsymbol{z}$ keeping $\boldsymbol{\beta}$ and $\boldsymbol{u}$ fixed. This is done independently for each covariate $1 \leqslant j \leqslant p$, because $R_1(\cdot)$ and the l_2^2-norm in (5) can be expressed as the sum of p terms each involving only the j-th row of $\boldsymbol{z} = (\boldsymbol{z}_0, \cdots, \boldsymbol{z}_K)$ and the j-th row of $\boldsymbol{\beta} + \boldsymbol{u}$. For example, when $a = 2$ in (4), we set $\boldsymbol{z}' = \Big(\boldsymbol{I} + \boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H}/\rho\Big)^{-1}(\boldsymbol{\beta} + \boldsymbol{u})'$, where the $K \times (K+1)$ matrix $\boldsymbol{H} = [-\mathbf{1}, \boldsymbol{I}]$ is the concatenation of $-\mathbf{1}$ and the K-dimensional identity matrix $\boldsymbol{I}$. This, computation is implemented by first computing the $(K+1) \times (K+1)$

matrix $\left(\boldsymbol{I} + \boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H}/\rho\right)^{-1}$, and then multiplying it with each column $j = 1, \cdots, p$ of $(\boldsymbol{\beta} + \boldsymbol{u})'$.

Lastly, in *step iii*, $\boldsymbol{u}$ is updated to $\boldsymbol{u} + \boldsymbol{\beta} - \boldsymbol{z}$. We iterate these three steps until the norm of both $\boldsymbol{z} - \boldsymbol{\beta}$ and the difference between $\boldsymbol{z}$ from two consecutive iterations becomes smaller than a pre-specified treshold $\epsilon > 0$ (Boyd et al., 2011).

4. Simulation Study

The ovarian cancer database *curatedOvarianData* (Ganzfried et al., 2013) includes 16 datasets (see Section 5). In our simulation study we considered a total of 18 studies. To evaluate variations in prediction accuracy we used either $K = 2, 5, 10$ or 15 of the 18 studies for model training. The remaining $16, 13, 8$ or 3 studies were used for external validations. For each study $k = 1, \ldots, 18$, we generated the sample size n_k from a uniform distribution $n_k \sim Unif(100, 101, \cdots, 500)$. In the *curatedOvarianData* library (Ganzfried et al., 2013) the sample sizes vary from small studies with less than 150 patients up to the 512 patients of the TCGA study. Individual covariates $\boldsymbol{x}_{k,i} \in \mathbb{R}^{500}, i = 1, \ldots, n_k$, have been generated from a normal distribution $\boldsymbol{x}_{k,i} \sim N_{500}(\boldsymbol{0}, \boldsymbol{V})$ with covariance $V_{j,j'} = \rho^{|j-j'|}$ (Tibshirani, 1997) and $\rho = 0.3$. Simulation scenarios with different values of ρ and different sample sizes n_k are considered in the supplementary material (Tables S2-S7).

We generated 100 times the parameters $\boldsymbol{\beta} \in \mathbb{R}^{500\times 19}$ and a collection of 18 studies $\mathcal{D} = (\mathcal{D}_k)_{k=1}^{18}$. In each of these 100 simulations we first generated the vector $\boldsymbol{\beta}_0 \in \mathbb{R}^{500}$ using a two-component mixture distribution with a point mass at zero and a Gaussian component with mean zero and variance σ_0^2. The proportion of zeros of this mixture is $p_0 = 0.9$ ($\sigma_0 = 0.5$) or 0 ($\sigma_0 = 0.1$). In the latter case $\beta_{0,j} \neq 0$ for every predictor $j = 1, \ldots, p$. In each simulation of the collection $\mathcal{D} = (\mathcal{D}_k)_{k=1}^{18}$ we independently generated $p = 500$ vectors $(\epsilon_{1,j}, \ldots, \epsilon_{K,j}) \sim N_K(\boldsymbol{0}, \boldsymbol{\Sigma})$ and set $\beta_{k,j} = \beta_{0,j} + \epsilon_{k,j}$ for each covariate $j = 1, \ldots, p$ and study $k = 1, \ldots, K$.

We consider three matrices $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3$ with 3, 2 or a single cluster of studies, see Figure 1. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

[Figure 1 about here.]

The individual outcomes were generated from proportional hazards models with regression coefficients $\boldsymbol{\beta}_k$, study-specific baseline survival functions $\widehat{S}_k(\cdot)$ and censoring distributions $\widehat{S}_{C,k}(\cdot)$. We estimated $\widehat{S}_k(\cdot)$ and $\widehat{S}_{C,k}(\cdot)$ from the *curatedOvarianData* library (Ganzfried et al., 2013). In our simulations the proportion of censored survival times ranged between 43% and 55% (see Supplementary Table S2). For each study k we also generated 1,000 additional observations, that were not used to fit regression models, but were used to evaluate predictions in population k.

4.1 *Selection of* Σ *and* (λ_1, λ_0)

We use initial estimates $\widehat{\boldsymbol{\beta}_k}$ obtained from K independent ridge regression models to estimate $\boldsymbol{\Sigma}$. The procedure leverages the Bayesian interpretation (2) of the regularized likelihood (1). As formalized in (2), with $R_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\boldsymbol{\beta}_{1:K,j} - \beta_{0,j}\mathbf{1}||_{\boldsymbol{\Sigma}}^a$, we can interpret, *a priori*, $(\beta_{j,1}, \cdots, \beta_{j,K}), j = 1, \cdots, p$, given $\boldsymbol{\beta}_0$, as p independent vectors. In particular, with $a = 2$, the parameters $(\beta_{j,1} - \beta_{j,0}, \cdots, \beta_{j,K} - \beta_{j,0}), j = 1, \cdots, p$, given $\boldsymbol{\beta}_0$, can be interpreted as independent multivariate normal vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. If $\boldsymbol{\beta}$ is known, we could therefore straightforwardly estimate $\boldsymbol{\Sigma}$.

Assuming $\lambda_0 = 0$, the joint normal distribution of $\boldsymbol{\beta}$ implies that $E[\boldsymbol{\beta}_k | \{\boldsymbol{\beta}_{k'}\}_{0<k'\leqslant K, k'\neq k}] = \sum_{0<k'\leqslant K, k'\neq k} \alpha_{k,k'}\boldsymbol{\beta}_{k'}$, where, for $k = 1, \cdots, K$, $\boldsymbol{\alpha}_k = (\alpha_{k,k'})_{0<k'\leqslant K, k'\neq k}$ is a function of $\boldsymbol{\Sigma}$. (Eaton, 1983) Therefore, the expectation of $\boldsymbol{X}_k\boldsymbol{\beta}_k$, conditional on $\{\boldsymbol{\beta}_{k'}\}_{0<k'\leqslant K, k'\neq k}$, is $\sum_{0<k'\leqslant K, k'\neq k} \alpha_{k,k'}\boldsymbol{X}_k\boldsymbol{\beta}_{k'}$. After replacing $\boldsymbol{\beta}_{k'}$ with $\widehat{\boldsymbol{\beta}}_{k'}$, we estimate $\boldsymbol{\alpha}_k$ via a Cox model with $K-1$ covariates $z_{k'} = \boldsymbol{X}_k\widehat{\boldsymbol{\beta}}_{k'}$. We then use $\boldsymbol{\beta}_k^\star = \sum_{0<k'\leqslant K, k'\neq k} \widehat{\alpha}_{k,k'}\widehat{\boldsymbol{\beta}}_{k'}, k = 1, \cdots, K$, and the resulting covariance matrix to select $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\beta}_k^\star$ has a direct interpretation under

the assumption that the vectors $(\beta_{1,j}, \ldots, \beta_{K,j})$ are in a linear subspace with dimensions less than K. Figure 2 shows averages across the 100 simulations of the estimated similarity matrix for the largest model with $K = 15$ studies when $p_0 = 0$ (top row) and $p_0 = 0.9$ (bottom row).

[Figure 2 about here.]

To select the parameters λ_0 and/or λ_1, we use Monte-Carlo cross-validation (CV). We compare candidate values (λ_0, λ_1) using the summary $\mathcal{C}(\widehat{\boldsymbol{\beta}}) = \sum_k w_k \mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$, where the C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k) = \widehat{Pr}\Big(\boldsymbol{x}_1'\widehat{\boldsymbol{\beta}}_k > \boldsymbol{x}_2'\widehat{\boldsymbol{\beta}}_k | Y_1 < Y_2\Big)$ estimates the concordance (Harrell Jr et al., 1984) between two independent and non-censored survival times Y_1 and Y_2 from the population k with risk scores $\boldsymbol{x}_1'\widehat{\boldsymbol{\beta}}_k$ and $\boldsymbol{x}_2'\widehat{\boldsymbol{\beta}}_k$. We first split each of the K datasets $\mathcal{D}_k$ randomly M-times into training (80%) and validation (20%) datasets. Next, we define a grid of tuning parameters. For each combination of the tuning parameters (λ_1, λ_0) we compute the estimates $\widehat{\boldsymbol{\beta}}^{(\ell)}$, $\ell = 1, \cdots, M$. We then select the (λ_1, λ_0) values with the highest average C-statistics.

4.2 *Prediction Accuracy*

Figure 3 shows, for each of the 18 studies box-plots of the estimated C-statistics (Harrell Jr et al., 1984) when either $2, 5, 10$ or 15 studies (1st to 4th column) were used to estimate the similarity matrix and the model (1). C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ for studies k that were utilized to estimate the model (estimated using the additional 1000 hold-out observations in study k) are highlighted inside the brown rectangles, whereas C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ for studies k that were not used to estimate the model are shown on the right of the brown rectangles.

The three rows of Figure 3 correspond to scenarios with data generated using $\boldsymbol{\Sigma}_1$ (top row of Figure 3), $\boldsymbol{\Sigma}_2$, (2nd row), or $\boldsymbol{\Sigma}_3$ (bottom row) as illustrated in Figure 1. Red, green and blue box-plots in the top-row indicate the three clusters of studies with $\boldsymbol{\Sigma}_1$. Similarly, red and green box-plots in the 2-nd row indicate the two clusters of studies with Σ_2. Differences

in the distribution of the C-statistics between studies within the same cluster are due to differences in the sizes n_k and covariate matrices $\boldsymbol{X}_k$, which remain identical across the simulated datasets.

[Figure 3 about here.]

With $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ (1st row of Figure 3) prediction accuracy improves when we increased the number of studies ($K = 2, 5, 10, 15$) used for model training. With $K = 2$ or 5, the studies used for model training belong to the first two clusters (red and green box-plots). In these two cases, for each study $k = 13, \cdots 18$ in cluster 3 (blue box-plots) the inter-quartile range (IQR) of the C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ across simulations lies within the interval 0.52-0.55. Whereas for $K = 10$ (3rd column, studies 1-4 and 10-15 were used for estimation), studies from all three clusters have been used for training. In this case, the IQRs of $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ across simulations for all three hold-out studies $k = 16, 17, 18$ in cluster 3 are within the interval 0.65-0.69. The last row of Figure 3 shows that, as expected, borrowing of information in the estimation of model parameters is most effective in the case of a single cluster of studies.

Next, we compared our estimates of $\boldsymbol{\beta}$ based on the hierarchical regularization (HR) model (1), with $a = 2$ for $R_1(\cdot)$ and ridge penalty (HR-R, $\lambda_1 = 0$) or LASSO penalty (HR-L, $\lambda_2 = 0$) for $\boldsymbol{\beta}_0$, to Cox models trained separately on each study $\mathcal{D}_k$ with LASSO (single-study LASSO, SL) or ridge penalties (single-study ridge, SR) for $\boldsymbol{\beta}_k$. In addition we consider two models that combine all (2, 5, 10 or 15) studies into a single dataset and estimate a single Cox model (with regression parameters $\boldsymbol{\beta}_0$) using a LASSO (pooled LASSO, PL) or ridge (pooled ridge, PR) penalty for the coefficients $\boldsymbol{\beta}_0$. We also consider two meta-analysis approaches described in (Waldron et al., 2014; Riester et al., 2014) that combine study specific estimates $\widehat{\boldsymbol{\beta}}_k$ into a single vector $\widehat{\boldsymbol{\beta}}_0$ using either fixed effects (FE) or random-effects (RE) estimation.

Figures 4, S3 and S4 show the average C-statistics of each method when we used $K = 5$ or $K = 10$ studies for estimation. The pooled LASSO and ridge models (PL and PR) and

the meta-analyses methods (FE and RE) estimate a single parameter $\boldsymbol{\beta}_0$, which was used to compute the C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ for each study k. For the single-study SL and SR models we used the study-specific estimates $\widehat{\boldsymbol{\beta}}_k$ to compute prediction $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ in population k (using the 1,000 validation observations). For prediction with SL and SR in populations k' not used for estimation, we used each estimate $\widehat{\boldsymbol{\beta}}_k$ of the $K = 5$ (or 10) training studies for predictions $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_{k'})$ in (*out of data collection)* populations k'. For each hold-out population k', we then averaged these $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_{k'})$ over all $K = 5$ (or 10) training studies, i.e. the rows for SL and SR in Figure 4 report $\sum_k \mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_{k'})/K$ for hold-out populations k'.

[Figure 4 about here.]

We use $\mathcal{K}$ to indicate the K populations represented in the collection of studies used for model training. For studies $k \in \mathcal{K}$ both HR-L and HR-R improve predictions $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ compared to single-study estimates SL and SR. For instance, with $K = 5, p_0 = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ (three clusters of studies), the average difference between $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ of HR-R and SR is between 0.03 and 0.08 for each of the five studies (0.73 to 0.78 for SR compared to 0.73 to 0.85 for HR-R). In contrast to HR-R and HR-L, Meta-analytic and pooled estimates (FE, RE, PR and PL) do not improve predictions in populations $k \in \mathcal{K}$ compared to SR.

For the HR, meta-analytic and pooling procedures we observe a higher average C-statistics in scenarios with $p_0 = 0.9$ ($\sigma_0 = 0.5$, 90% of the components of $\boldsymbol{\beta}_0$ are null) compared to scenarios with $p_0 = 0$ ($\sigma_0 = 0.1$, $\beta_{0,j} \neq 0$ for all j). For example, when we consider $K = 10$, $p_0 = 0$ and HR-L predictions in populations $k \notin \mathcal{K}$ not used for model training, the average C-statistics across simulations was between 0.69 and 0.7. Instead with $p_0 = 0.9$ the average C-statistics of HR-L varied between 0.74 and 0.76 (an increase between 0.04 and 0.07). For meta-analytic (FE, RE) and pooling (PL, PR) procedures we observed similar differences when we compare simulation results with $p_0 = 0.9$ and $p_0 = 0$.

The meta-analytic (RE and FE) and pooling (PL and PR) methods provide a single

prediction model, developed assuming identical (FE, PL and PR) or exchangeable (RE) covariate effects across studies. These methods estimate average covariate effects across studies. The HR-L and HR-R procedures generate K prediction models, with parameters $\widehat{\boldsymbol{\beta}_k}$ for populations $k \in \mathcal{K}$, and an additional $(K+1)$-th prediction model with parameter $\widehat{\boldsymbol{\beta}_0}$ applicable in populations $k' \notin \mathcal{K}$ beyond the K training populations. In our simulation study (scenarios $\Sigma = \Sigma_3$) we note similar prediction accuracy (C-statistics) in external validation datasets $k' \notin \mathcal{K}$ across methods (HR-R, HR-L, RE, FE, PR and PL). In contrast, the left columns of Figure 4 shows validation summaries based on hold-out samples for populations $k \in \mathcal{K}$. These columns illustrate, in scenarios with different covariate-outcome relations in 3 clusters of studies ($\Sigma = \Sigma_1$, Figure 1), that prediction models tailored to the populations $k \in \mathcal{K}$ tend to achieve better accuracy summaries compared to RE,FE, PL and PR.

With a single cluster of studies ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_3$, see Figures 1 and S4, and similar parameters $\boldsymbol{\beta}_k$ across studies, the scenario becomes ideal for the application of pooling procedures. PL, PR, HR-R and HR-L predict survival substantially better than the FE, RE, SL and SR methods (Figure S4), With $p_0 = 0$, predictions based on the HR-R and HR-L procedures in hold-out samples of populations k used for model training are on average better than for PR and PL (difference of 0.01 to 0.04 for HR-R compared to PR with $K = 5$ studies, and 0.02 to 0.05 with $K = 10$). We also observe minimal improvements of the HR-R and HR-L procedures compared to the PR and PL procedures when we consider the C-statistics for predictions in the holdout studies $k \notin \mathcal{K}$.

We also examined additional simulation scenarios (Supplementary Material) with different sample sizes and covariate-outcome relations across studies $k \in \mathcal{K}$. In particular, Table S6 considers data collections with highly unbalanced sample sizes n_k across clusters of studies. Whereas Table S7 considers data collections that include one outlier study, with markedly discordant covariate-outcome relations compared to the other studies.

5. Survival prediction in ovarian cancer

We applied model (1) to predict survival in ovarian cancer using the *curatedOvarianData* database (Ganzfried et al., 2013), a curated collection of gene-expression datasets. To evaluate prediction, we split the largest study in the database, the TCGA dataset with 510 observations, 100 times randomly into a training dataset of $n_1 = 50, 75, \ldots,$ or 300 observations and a validation dataset with $510 - n_1$ observations. We predicted patient survival $Y_{1,i}$ in the TCGA holdout data using five additional datasets (PMID-17290060, GSE51088, MTAB386, GSE13876, GSE19829) with sample sizes ranging between 42 (GSE19829) and 157 (GSE13876) observations. The proportions of censored survival times in the TCGA, PMID-17290060, GSE51088, MTAB386, GSE13876 and GSE19829 datasets were 42%, 42%, 24%, 43%, 28% and 45%. For external validations of model (1) we used two large datasets (GSE9891, $n_k = 258$ and GSE26712, $n_k = 185$) that were not used to train the model. In all the analyses we used the expression values of the $p = 3,030$ genes that are common in all studies to predict survival.

To evaluate our model (1), we created different cross-study heterogeneity scenarios that are motivated by documented inconsistencies across cancer datasets and by possible preprocessing errors (Potti et al., 2006). This was achieved by introducing in one (scenario 2: GSE13876) or two studies (scenario 3: GSE13876 and GSE19829) a distortion of the expression values $x_{k,i,j}$ which become $10 - 3x_{k,i,j}, j = 1, \ldots, p$ for study $k = K$ (scenario 2) or studies $k = K - 1$ and K (scenario 3). Whereas in scenario 1 we used the covariates $x_{k,i,j}$ of the six studies.

Similar to Section 4, we considered parameter estimates based on the $n_1 = 50, \cdots, 300$ TCGA training samples using (i) single study Cox models with LASSO (SL) or (ii) ridge (SR) regularization, pooled Cox regression models that combine the n_1 TCGA-observations and the remaining five studies (PMID-17290060, GSE51088, MTAB386, GSE13876 and

GSE19829) into a single dataset with (iii) LASSO (PL) or (iv) ridge (PR) regularization, (v) fixed effects (FE) and (vi) random effects (RE) meta-analyses models as described in Riester et al. (2014); Bernau et al. (2014); Waldron et al. (2014), and (vii) the proposed hierarchical regularization model (1) with $\lambda_0 = 0$ and $a = 2$ (HR-R).

Single-study Cox models with LASSO-penalty trained on the TCGA data with $n_1 = 50$ data points had a low average C-statistics of 0.51 across the 1,000 training-validation partitions, with minor improvements up 0.52 when $n_1 = 300$ observations were used for model training. Single study ridge regression models performed substantially better, with average C-statistics ranging between 0.53 for $n_1 = 50$ and 0.58 for $n_1 = 300$ observations (Figure S5). Improvements through integration of additional studies varied substantially across data-integration methods. For scenario 1, FE and RE meta-analyses have both nearly constant and identical average C-statistics of 0.57 across all sample sizes n_1. Whereas PR and PL had an average C-statistics of 0.56 and 0.60 for $n_1 = 50$ with improvements up to 0.57 and 0.61 when $n_1 = 300$. HR-R had an average C-statistics of 0.61 when $n_1 = 50$ and 0.62 when $n_1 = 300$.

Figure 5 shows, for scenarios 2 and 3, average C-statistics for the TCGA validation samples. Different curves correspond to different prediction methods. The black curves show the average C-statistics (y-axis) across 100 TCGA validation subsets (size $510 - n_1$) with Cox models trained on $n_1 = 50, \cdots, 300$ observations from the TCGA study (x-axis) using either SL (dotted curve) or SR (solid curve). The red curves show the average C-statistics for PR (solid curve) and PL (dotted curve) models, the green curves correspond to FE (dotted line) and RE (solid line) meta-analysis models, and the blue curve corresponds to the HR-R.

[Figure 5 about here.]

In scenario 2, the RE meta-analysis, which combines estimates from $n_1 = 50$ TCGA data points with estimates from the remaining five studies, has the same average C-statistics

as the single-study SR model trained on $n_1 = 240$ patients. For sample sizes $n_1 > 275$, the pooled regression models PR and RE have similar performances. The HR-R model trained on $n_1 = 50$ TCGA patients has an average C-statistics of 0.61 and was superior to the remaining procedures. As expected, with additional discrepancies in the relations between covariates and outcomes across studies (scenario three), the performances of all data-integration methods decrease. The HR-R model had an average C-values of 0.60 with $n_1 = 50$ TCGA patients which increases up to 0.61 when $n_1 = 300$ PR, PL, FE and RE methods rely on the assumption that the regression parameters are similar across studies. With substantial departures from this assumption the hierarchical model HR-R shows, across all sample sizes $50 \leqslant n_1 \leqslant 300$, substantial gains in prediction accuracy compared to PR, PL, FE and RE.

In consideration of the results in Figure 5, we tested if the prediction accuracy (C-statistics) with HR-R outperformed meta-analysis (RE). We tested the null hypothesis H_0 : *the HR-R prediction model has a lower or equal accuracy (concordance index in the TCGA population) than the RE model*, at significance level $\alpha = 0.05$. We generated $c = 1, \cdots, 20$ independent partitions of the TCGA dataset into training ($n_1 = 150$ patients) and a validation (360 patients) components. For each partition $c \leqslant 20$ we compared the prediction models (C-statistics for HR-R and RE) using the TCGA validation component. Then, for each partition we used a standard bootstrap algorithm to test H_0, resampling 999 times the TCGA validation component. In scenario 2 (scenario 3) H_0 was rejected 15 (16) times out of the 20 partitions.

Supplementary Figure S6 shows for Scenario 1 the C-statistics in external validations with datasets GSE9891 and GSE26712, which were not used for model training. Model training involved n_1 TCGA samples combined with five additional studies (PMID-17290060, GSE51088, MTAB386, GSE13876 and GSE19829). In the GSE26712 validation dataset the

C-statistics with HR-R increased from 0.64 for $n_1 = 50$ up to 0.66 when $n_1 = 300$, while meta-analytic and pooling procedures had a C-statistics between 0.55 and 0.58 (0.55-0.56 for PR, 0.55-0.58 for PL, 0.56-0.58 for FE and 0.56-0.57 for RE). For GSE26712, we observe modest improvements in prediction accuracy of HR-R compared to meta-analytic and pooling procedures.

Supplementary Table S8 reports the C-statistics in the TCGA hold-out data for an extension of our HR-R procedure which combines gene expression profiles with additional demographic and clinical covariates (age at diagnosis, tumor stage and grade, and proportions of tumor and stromal cells in the biopsy). These 5 covariates are available in the TCGA dataset, but where not available in all datasets. We observed minor variations in prediction accuracy (C-statistics) of the HR-R procedure by including these additional covariates.

6. Discussion

The joint analysis of omics variables and time-to-event outcomes, and the use of individual profiles $\boldsymbol{x}_{k,i}$ for predictions, are particularly challenging when the sample size is small. These analyses often include thousands of potential predictors. The use of multiple studies and pooling of information can improve prediction accuracy. Meta-analyses can be utilized when the relations of covariates and outcomes are homogeneous across studies. But recent work in oncology (Riester et al., 2014; Waldron et al., 2014; Trippa et al., 2015) showed that there can be clusters of studies with relevant discrepancies in their covariate-outcome relations.

Ren et al. (2020) introduced *specialist* and *generalist* predictions. *Specialist prediction models* tailor predictions to a population of interest $k \in \{1, \ldots, K\}$ (e.g., patients diagnosed at a specific institution k, or more generally a well defined group of patients) when the data collection includes samples $\mathcal{D}_k$ from the population k. In this manuscript we discussed a method to leverage additional datasets $\mathcal{D}_{k'}$, beyond dataset $\mathcal{D}_k$, accounting for potential differences between the covariate-outcome relations in K distinct populations. *Generalist*

prediction models consider a different prediction problem. Predictions are still based on a collection of K datasets, representative of distinct populations (e.g. institutions that adopted different protocols or assays for data collection). The joint distributions of covariates and outcomes can vary across populations. The primary aim in this case is to predict in contexts beyond the K populations that were sampled (e.g. a prediction model for institutions that were not included in the data collection).

Ren et al. (2020); Bernau et al. (2014); Trippa et al. (2015) showed that standard cross validation procedures are adequate to evaluate specialist prediction models, while more stringent validation schemes are necessary to evaluate generalist prediction models. Specialist and generalist models have different applications, and in most cases they are not interchangeable. For example, specialist models tailored to specific groups of patients (e.g. patients receiving distinct treatments or in different geographical regions) can support clinical decisions (Chen et al., 2009). Borrowing information across groups (e.g. distinct treatments or institutions) can improve their prediction accuracy. In other applications the development of a generalist model is crucial, for example to develop prevention policies (Evans and Howell, 2007).

Careful selection of covariates and data pre-processing (e.g.,imputation of missing data, normalization, expression levels in tumor and stromal cells, etc.) is an important component of the analysis. Our model is applicable to any set of post-processed predictors $\boldsymbol{x}_{k,i}$. Selected prognostic variables j could be included without regularization of the corresponding parameters $\beta_{0,j}$. Indeed one can modify $R_0(\cdot)$ and include covariate-specific penalties $R_0(\boldsymbol{\beta}_0) = \sum_j \lambda_{0,j}|\beta_{0,j}| + \frac{\lambda_{1,j}}{2}\beta_{0,j}^2$, and select (i) $\lambda_{0,j} = \lambda_{1,j} = 0$ for selected prognostic variables and (ii) positive parameter values (λ_0, λ_1) for the remaining covariates. Alternatively one can compute preliminary estimates $\widehat{\beta}_{0,j}$ and use $\lambda_{\ell,j} = \lambda_\ell / \widehat{\beta}_{0,j}^{\gamma}$, $\gamma > 0$ (Zou, 2006).

Our real-data applications in Section 5 only considered covariates (gene expression) available in all six studies to estimate the prediction models. We conducted additional analysis

using data imputations to relax this constraint. We considered prediction in the TCGA hold-out data ($n_1 = 150$) based on single-study, meta-analytic, pooling and HR-R procedures. For model training, we either used genes included in all six datasets or all genes available in the TCGA dataset. For the methods that we considered (SR, RE, FE, PR and HR-R), the extended list of covariates did not improve prediction accuracy (metric: C-statistics).

We combined two established concepts, regularization of regression models (Hoerl and Kennard, 1970; Tibshirani, 1997) and metrics of similarity between datasets that identify clusters of studies. We used these concepts to estimate study-specific parameters $\boldsymbol{\beta}_k$ and for prediction, both in the K populations represented in the data collection, and in other contexts by estimating the parameter $\boldsymbol{\beta}_0$.

The $K \times K$ similarity matrix $\boldsymbol{\Sigma}$ is used to regularize the likelihood function, and it tunes the degree of borrowing of information in the estimation of K study-specific regression models. It shrinks the estimate of the k-th study-specific regression parameter $\boldsymbol{\beta}_k$ towards $\boldsymbol{\beta}_{k'}$ when the studies k and k' are similar. In contrast studies k' with low similarity ($\Sigma_{k,k'} \approx 0$) have little influence on the estimation of $\boldsymbol{\beta}_k$. In our analyses we illustrated that, if the data collection includes clusters of studies with similar predictors-outcome relations, then the introduced method improves the prediction accuracy compared to alternative procedures, including single-study estimates, meta-analyses and pooling all K studies into a single data matrix.

Acknowledgements

SV and LT were partially supported by NIH Grant 1R01LM013352-01A1. RM was partially supported by ONR Grants ONR-N000141512342, ONR-N000141812298, and NSF Grant NSF-IIS-1718258.

Data Availability Statement

The data that support the findings in this paper are openly available in the Bioconductor package *curatedOvarianData* (Ganzfried et al., 2013) at `http://doi.org/10.1093/database/bat013`.

References

Bernau, C., Riester, M., Boulesteix, et al. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30,** i105–i112.

Boyd, S., Parikh, N., Chu, E., Peleato, and et al. (2011). Distributed optimization and statistical learning via the admm. *Foundations and Trends in Machine Learning* **3,** 1–122.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *J on Scientific Com* **16,** 1190–1208.

Chen, S., Blackford, A. L., and Parmigiani, G. (2009). Tailoring brcapro to asian-americans. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **27,** 642.

Cheng, X., Lu, W., and Liu, M. (2015). Identification of homogeneous and heterogeneous variables in pooled cohort studies. *Biometrics* **71,** 397–403.

Conlon, E., Postier, B., Methe, B., and et all (2009). Hierarchical bayesian meta-analysis models for cross-platform microarray studies. *J Applied Stat* **36,** 1067–1085.

Conlon, E. M., Postier, B. L., Methé, B. A., Nevin, K. P., and Lovley, D. R. (2012). A bayesian model for pooling gene expression studies that incorporates co-regulation information. *PloS one* **7,** e52137.

Conlon, E. M., Song, J. J., and Liu, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics* **7,** 1.

Cox, D. R. (1972). Regression models and life-tables. *JRSS(B)* **34,** 187–220.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials* **7,** 177–188.

Eaton, M. L. (1983). Multivariate statistics: a vector space approach. *WILEY* .

Evans, D. G. R. and Howell, A. (2007). Breast cancer risk-assessment models. *Breast cancer research* **9,** 1–8.

Feng, F. Y., Huang, H.-C., Spratt, D. E., Zhao, S. G., Sandler, H. M., Simko, J. P., Davicioni, E., Nguyen, P. L., Pollack, A., Efstathiou, J. A., et al. (2021). Validation of a 22-gene genomic classifier in patients with recurrent prostate cancer: An ancillary study of the nrg/rtog 9601 randomized clinical trial. *JAMA oncology* .

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Comp & Graphical Stat* **7,** 397–416.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., et al. (2013). curatedovariandata: clinically annotated data for the ovarian cancer transcriptome. *Database* **2013,**.

Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modeling strategies for improved prognostic prediction. *Stat Med* **3,** 143–152.

Hedges, L. V. and Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Liu, F., Dunson, D., and Zou, F. (2011). High-dimensional variable selection in meta-analysis for censored data. *Biometrics* **67,** 504–512.

Liu, M., Lu, W., Krogh, V., Hallmans, G., Clendenen, T. V., and Zeleniuch-Jacquotte, A. (2013). Estimation and selection of complex covariate effects in pooled nested case–control studies with heterogeneity. *Biostatistics* **14,** 682–694.

Ma, S., Huang, J., and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* **12,** 763–775.

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., et al. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nature medicine* **12,** 1294–1300.

Ren, B., Patil, P., Dominici, F., Parmigiani, G., and Trippa, L. (2020). Cross-study learning for generalist and specialist predictions. *arXiv preprint arXiv:2007.12807* .

Riester, M., Wei, W., Waldron, L., Culhane, A. C., et al. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *JNCI* page dju048.

Sinha, D., Ibrahim, J. G., and Chen, M. (2003). A bayesian justification of cox's partial likelihood. *Biometrika* **90,** 629–641.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat Med* **16,** 385–395.

Trippa, L., Waldron, L., Huttenhower, C., and Parmigiani, G. (2015). Bayesian nonparametric cross-study validation of prediction methods. *Ann. Appl. Stat.* **9,** 402–428.

Waldron, L., Haibe-Kains, B., Culhane, A. C., et al. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *JNCI* **106,** dju049.

Wei, L. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat Med* **11,** 1871–1879.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101,** 1418–1429.

Supporting Information

Web Appendices S2 and S3 referenced in Section 4, Web Appendix S4 referenced in Section 5, and JULIA code implementing the HR-R and HR-L models, are available with this paper at the Biometrics website on Wiley Online Library

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

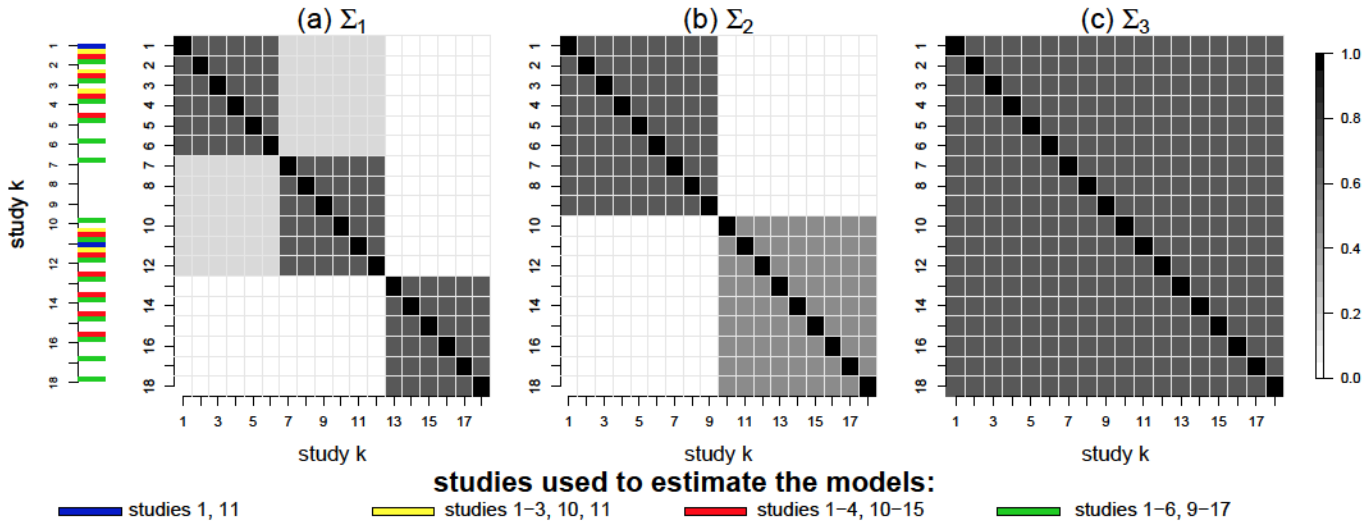

Figure 1. Similarity matrices $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3 \in \mathbb{R}^{18\times 18}$ used to simulate the datasets. We simulated collections of 18 datasets $\mathcal{D} = \{D_k\}_{k=1}^{18}$ with similarity matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{\beta}$. Studies indicated in blue (2 studies), yellow (5 studies), red (10 studies) or green (15 studies) are used to fit models with $K = 2, 5, 10$ or 15 studies. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

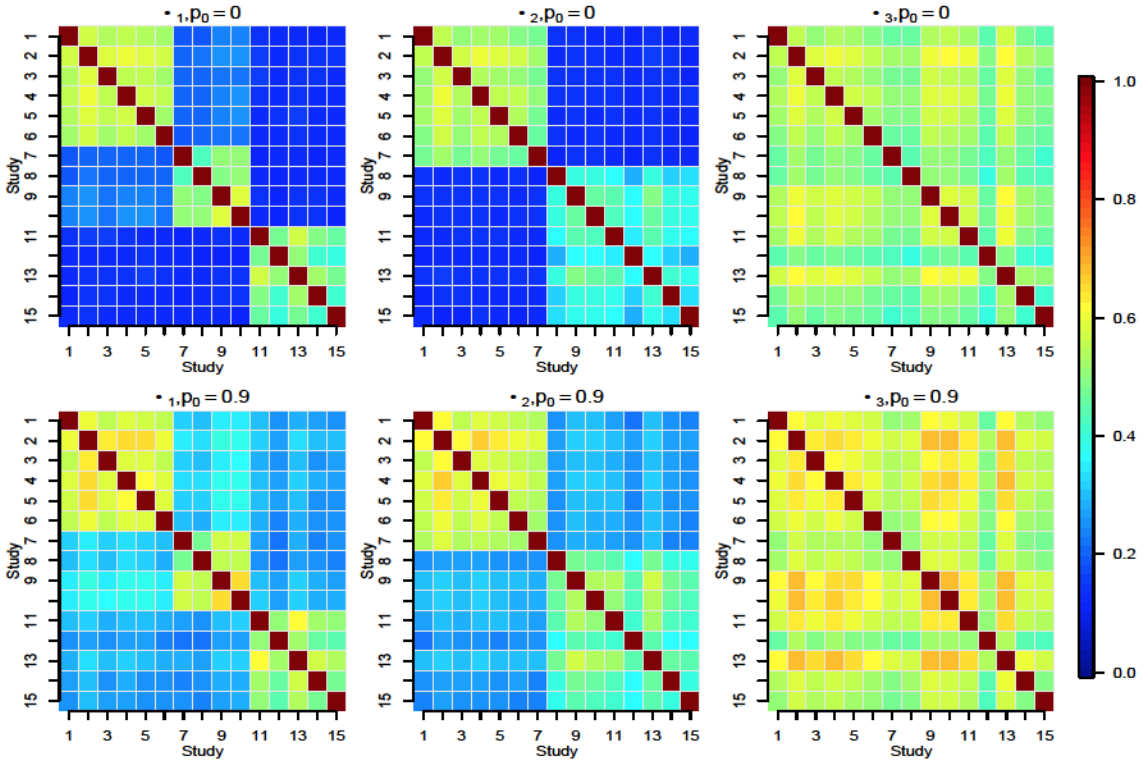

Figure 2. Average estimates across 100 simulations of the similarity matrix with $K = 15$ studies (studies 1-6 and 9-17 indicated in green in Figure 1). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

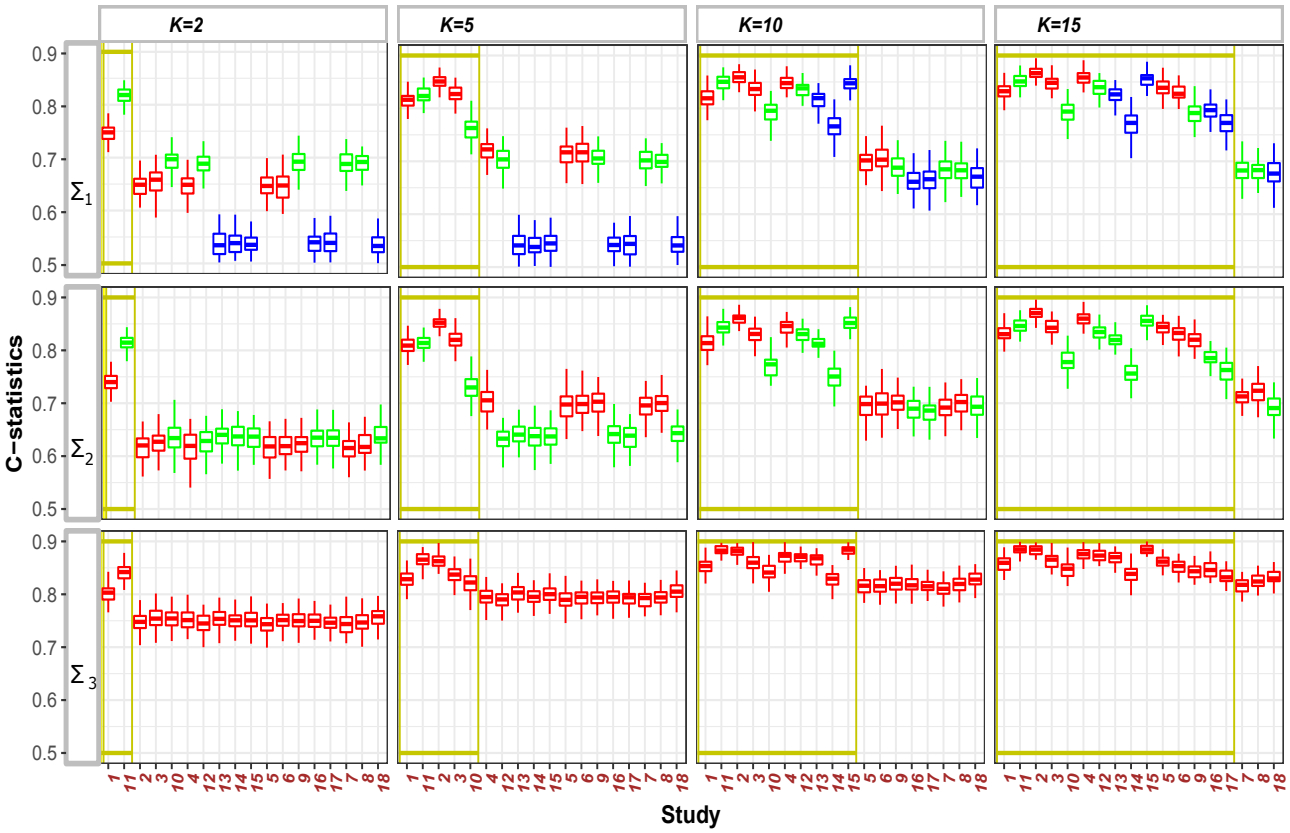

Figure 3. Predictions with the penalized regression model (1) for $(a, \lambda_0) = (2, 0)$. We consider $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ (see Figure 1), $p_0 = 0$ and 100 simulations of a collection of 18 studies. Either $K = 2, 5, 10$ or 15 studies (studies inside the brown rectangles) were used for selection/estimation of $(\boldsymbol{\Sigma}, \lambda_1, \boldsymbol{\beta})$. The colors (red, green, blue) of the Box-plots indicate clusters of studies (3, 2, or 1 clusters when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ or $\boldsymbol{\Sigma}_3$). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

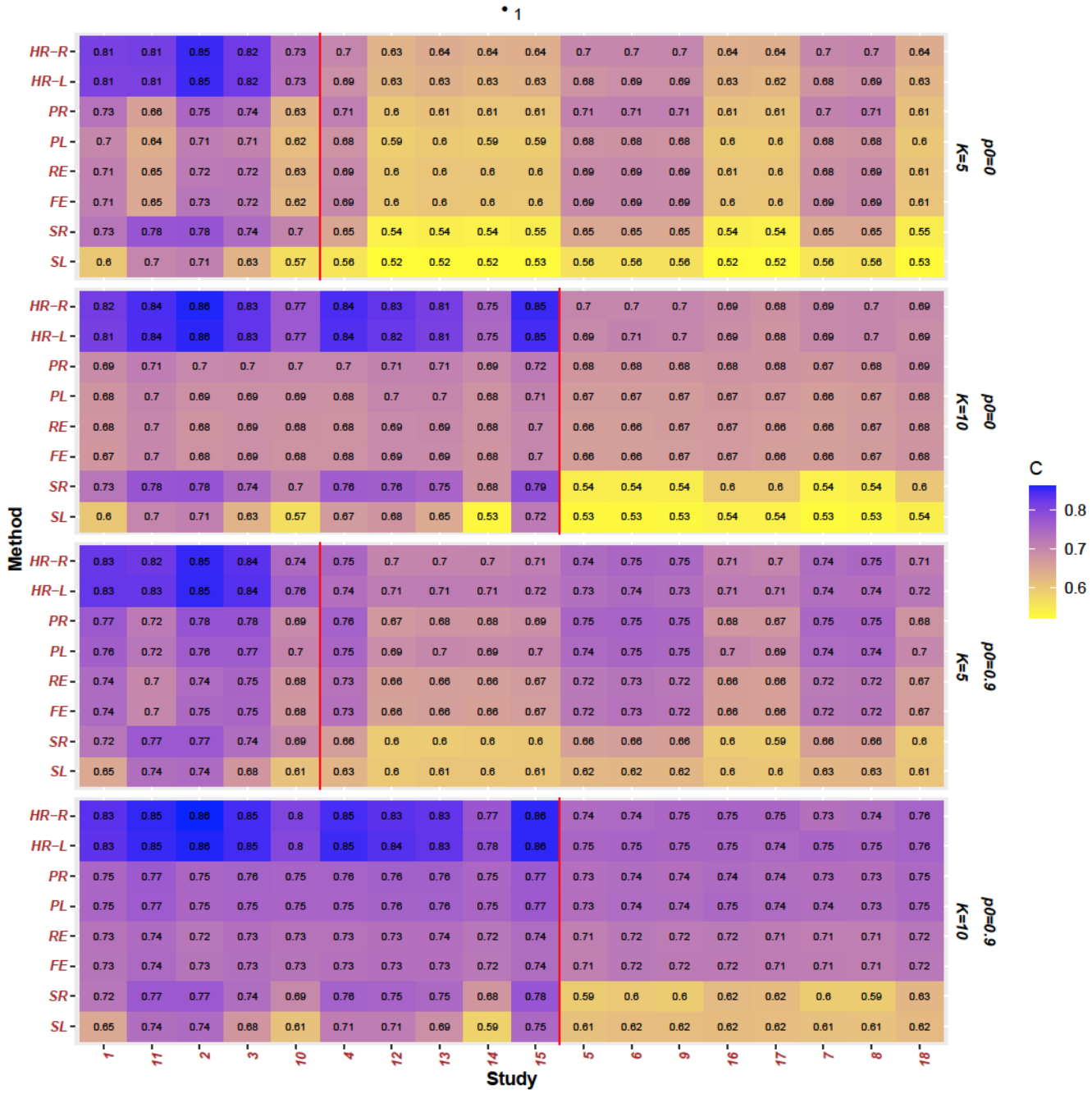

Figure 4. Average prediction accuracy (C-statistics) across 100 simulations of a collection of 18 studies. Study specific effects β_k have been generated under Σ_1 (see Figure 1). Either 5 or 10 of the 18 studies (studies on the left of the vertical red bar) are used to estimate the similarity matrix and covariate effects. See Figures S3 and S4 for results with similarity matrices Σ_2 and Σ_3. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

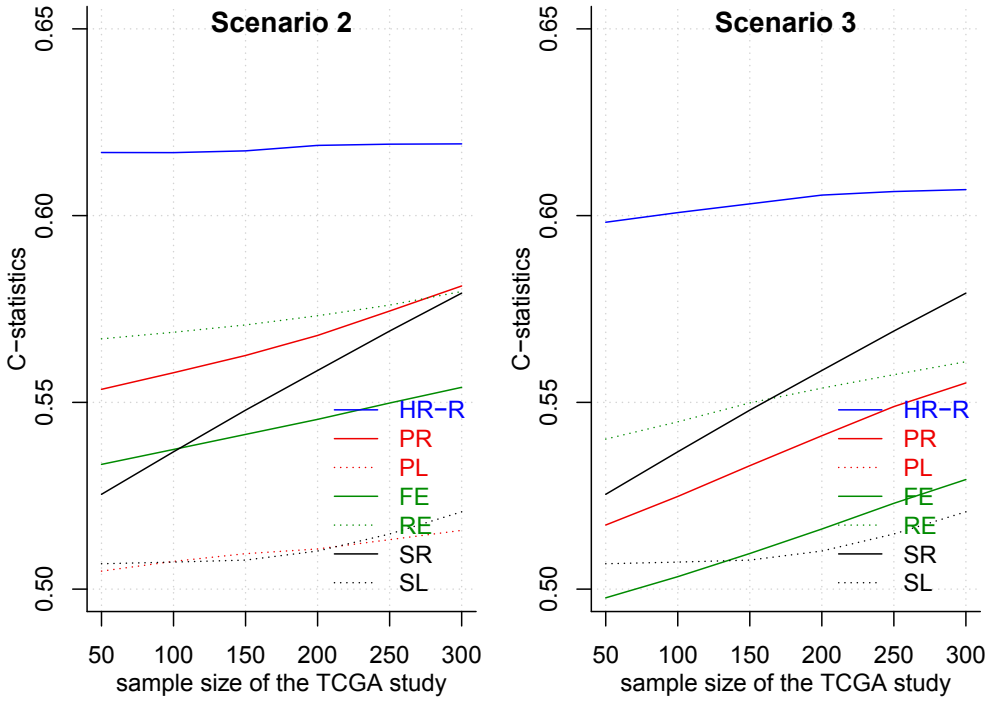

Figure 5. Average C-statistics for single-study SL and SR methods with $n_1 = 50, \cdots, 300$ TCGA training samples, and for data-integration methods (PL, PR, FE, RE, HR-R), which used n_1 TCGA training samples and five additional studies (PMID-17290060, GSE51088, MTAB386, GSE13876 and GSE19829) for model training. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.